EVWSD-ITA at EVALITA 2026: Overview of the Enhanced Visual Word Sense Disambiguation for Italian Task

Elio Musacchio[1,2,*], Lucia Siciliani[1], Pierpaolo Basile[1] and Giovanni Semeraro[1]

[1]*Dept. of Computer Science, University of Bari Aldo Moro, Via E. Orabona, 4 - 70125 Bari (ITALY)*
[2]*Dept. of Computer Science, University of Pisa, Largo Bruno Pontecorvo, 3, Pisa (ITALY)*

Abstract

EVWSD-ITA (Enhanced Visual Word Sense Disambiguation for Italian) is a shared task proposed at the EVALITA 2026 campaign. While traditional Visual Word Sense Disambiguation (VWSD) focuses on broad semantic distinctions, EVWSD-ITA introduces a more rigorous challenge by requiring systems to perform fine-grained disambiguation. The task involves selecting the most appropriate image for a target word within a given context, specifically designed to include "hard negatives"—co-hyponyms that share a common hypernym but represent distinct concepts. The dataset was meticulously constructed, comprising a manually validated test set. In this report, we showcase the dataset construction procedure and the results of the task.

1. Introduction

Word Sense Disambiguation (WSD) is a historical task in the Natural Language Processing field of research [1]. In this task, the objective is to select the correct sense of a target word in a sentence, out of all its possible meanings. Given its longevity, several works have explored algorithms and neural networks capable of solving the task [2]. Nevertheless, this task still cannot be solved entirely, thus capturing all the nuances of natural language remains an intriguing challenge. The introduction of Large Language Models (LLMs) has further ignited interest in resolution of this task, with several works proposing usage of LLMs to perform WSD [3]. In light of the success and interest of WSD, Visual Word Sense Disambiguation (VWSD) [4] has been proposed. In this task, the objective is to select an image, out of ten possible candidates, which correctly represents the sense of a target word in a given sentence. The sentence consists of a word of interest (that is, the target to disambiguate) and additional context words to support disambiguation (e.g., "music keyboard", where "keyboard" is the target to disambiguate and "music" is the context word). The gold standard is represented by examples annotated manually, where human experts have selected images that are meaningful with respect to the text, exploiting BabelDomains from BabelNet. Furthermore, the original VWSD dataset is split into three languages: English, Italian and Farsi. However, the task has a very clear limitation: it does not consider fine-grained semantics. Considering the previous example, for the target word "music keyboard", any image not related to music can be discarded, which favors systems with a good understanding of high-level semantic relationships across language and vision. The task does not include recognition of specific senses (e.g., disambiguating the image of a piano from the image of a harpsichord). In light of this, we propose a new task that combines both high-level and fine-grained semantics. The goal is not only to identify the broad sense of the target word, but also to accurately recognise its specific sense. In light of this, we propose the Enhanced Visual Word Sense Disambiguation for Italian task (EVWSD-ITA) at EVALITA 2026 [5].

The proposed task is significantly different from others in the state-of-the-art for the following reasons:

- We incorporate both high-level and fine-grained semantics by combining images representing different senses of the target word and co-Hyponyms (synsets that share the same Hypernym)
- We focus on the Italian language and manually annotate the test set to guarantee the quality and rigorousness of the benchmark

2. Related Works

The original WSD task has been extensively studied in previous research. In this context, one of the most used datasets is XL-WSD [6], a cross-lingual evaluation benchmark for the WSD task featuring sense-annotated development and test sets in 18 languages from six different linguistic families. Originally, WSD disambiguation models leveraged the BERT model. For example, Huang et al. [7] proposed GlossBERT, a family of models fine-tuned on context-gloss pairs and then converting WSD to a sentence-pair classification task. More recent approaches started to focus on leveraging LLMs for WSD. For example, Yae et al. [8] proposed three techniques to leverage LLMs in WSD. They format the task in the following formats: 1) multiple-choice questions; 2) binary questions; 3) multiple-choice questions for unseen words. Furthermore, Meconi et al. [9] propose an extensive study on the capabilities of LLMs in the WSD task, with a focus on the lexical understanding capabilities of the best performing models.

Following the success of WSD, the VWSD task was proposed for SemEval 2023 and, even after the completion of the challenge, several solutions and extensions of the dataset were proposed. Kritharoula et al. [10] tested several approaches using both generative LLMs for phrase enhancement and encoder-based vision-language models for retrieval. Specifically, the authors explore several possible solutions to address the VWSD task, they also consider image captioning and image retrieval, in order to cast the task as a unimodal retrieval task, as well as a learning to rank system and question-answering with chain-of-thought prompting. Kwon et al. [11] proposed an unsupervised VWSD approach exploiting definition generation with GPT-3 and Bayesian inference. Specifically, the authors perform context-aware definition generation with GPT-3 to overcome the out of vocabulary problem (that is, not all words having their definitions available in a lexical knowledge-base). Then they perform Bayesian style inference for Image-Text Matching. Zhang et al. [12] proposed SRCB, which employs a three stage pipeline: a context retrieval module, which predicts the correct definition using a bi-encoder architecture, a image retrieval module, which retrieves the relevant images from an image dataset, and a matching module, which decides to use either text or images and ranks them. Yang et al. [13] proposed MTA, which employs self-distillation to align fine-grained textual features to fixed vision features and align non-English textual features to English textual momentum features. Furthermore, they introduced a trilingual image-text dataset for VWSD, encompassing a fine-grained network of 85,754 word-sense associations and 120,131 images. Laba et al. [14] extended the original dataset to include the Ukrainian language. The authors extended the dataset through a semi-supervised approach leveraging Wikipedia articles and expert annotators. Furthermore, they test the performance of eight multilingual and multimodal LLMs using this dataset. Setitra et al. [15] generated visually representative images from textual descriptions, as well as rich textual descriptions from images. Then, an ensemble of deep models is used to perform classification. Musacchio et al. [16] proposed to use strictly hard negative samples exploiting the co-Hyponym relation from semantic networks. The authors then propose a new benchmark on a dataset created using this strategy and extensively evaluate both Vision-Language Encoders and Large Language Models supporting multimodal inputs.

Despite the research interest and popularity of the task, no work considers the proposal of a benchmark that combines high-level and fine-grained semantics on this task.

3. Task Description

We propose a single task, that is Visual Word Sense Disambiguation. Given a query and ten possible candidate images for it, we want to learn a system that is capable of selecting the relevant images for the query. In the task, only one candidate image is relevant for the query. Hence, the system must be able to either: 1) predict the one relevant image; 2) provide a score for each image based on its relevance to the query.

In EVWSD-ITA, we combine two different types of images: 1) images representing a different sense of the target word, for example the word "keyboard" may be paired to both the image of a computer keyboard and the image of a piano; 2) images representing a different specific sense for the target word while sharing the same broad sense, for example the word "keyboard" may be paired to both the image of a piano and the image of a harpsichord. We illustrate this in Figure 1. We use co-Hyponyms extracted from a semantic network to find *hard negatives* [17, 16]. That is, the images are related to the general sense of the target word, but that represent a different specific sense. We propose to improve the task by mixing the two types of images: 1) images for other senses of the target word; 2) images that share the same broad sense as the target word. This will test the ability of the proposed solutions to understand both fine-grained and high-level semantics.

In the original VWSD challenge, two metrics were used to assess the ability of the proposed solutions: HIT@1 and MRR (Mean Reciprocal Rank). It may be impossible to compute MRR in some cases, since it expects a ranking (e.g., systems trained for prediction only rather than ranking cannot be benchmarked using MRR). Hence, we will consider two different leaderboards: one focusing on MRR and the other focusing on HIT@1. Thanks to this, we can distinguish the goodness of the proposed solution on two different aspects (precise classification and overall ranking quality). We will describe the metrics in detail in Section 5.1.

4. Dataset

In order to propose the EVWSD-ITA task, we need a dataset where images satisfy the properties we are interested in (images representing synsets from that co-Hyponym of the target and synsets that share the same lemma as the target).
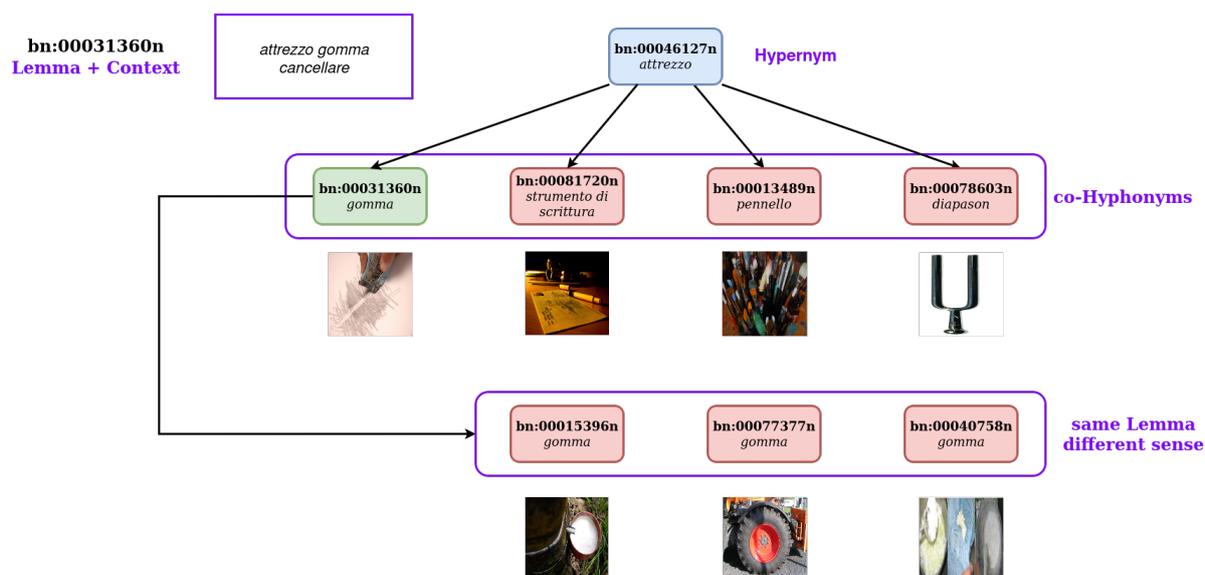

Figure 1: Example of the proposed approach. The synset in green is the correct one, while the ones in red are negatives.

This section describes the creation of the dataset and provides quantitative statistics.

4.1. Dataset Construction

Our data collection relies on BabelNet [18]. Since it provides images for each synset, it is a valuable resource to obtain data for our task. For each synset in BabelNet with an English gloss, we extracted the synset identifier and the corresponding Italian gloss. Since a synset can have multiple glosses, we select only one gloss according to its source. In particular, for selecting only one gloss, we prioritised each source according to this rank: WordNet, Wiktionary, Wikipedia, Wikidata, and other sources. Then, the URL's image from BabelNet is extracted for each synset. Since more images can be assigned to each synset, we used the image tagged with the attribute "best image". Each synset $s_i \in S$, where S is the set containing all synsets in BabelNet, has multiple lemmas $L(s_i) = \{l_1, ..., l_{m_i}\}$, where m_i is the number of lemmas for synset s_i, and a gloss g_i associated to it. We define a relationship $Hyp : S \rightarrow S$ which connects each synset to its Hyponym. Furthermore, we define two sets associated to s_i, the co-Hyponyms set $CH(s_i) = \{s_j \mid Hyp(s_j) = Hyp(s_i) \wedge s_j \neq s_i \wedge s_j \in S\}$ and the set of synsets with the same lemma $SL(s_i, l_i) = \{s_j \mid \exists l_j \in L(s_j) : l_j = l_i \wedge s_j \neq s_i \wedge s_j \in S\}$.

For the test set, we created the data using the following methodology: we iterated over all lemmas in the collection and iterated over all possible synsets for each lemma. We discarded instances where there were no synsets in the co-Hyponyms set or in the synsets with the same lemma. Then, we manually checked each instance and the candidate images of each synset. We removed synsets from the candidates when: 1) the distinct synset specific sense was too similar to that of the target sense; 2) visual elements of the image associated to the synset were not enough to distinguish it with respect to the target synset. Specifically, for the first case, we found some synsets in BabelNet that shared the same gloss or represented the same sense as the target synset but with a different Hyponym. For the second case, we found cases where distinct synsets were represented using the same image or there were not enough visual elements to distinguish it with respect to the image of the target synset (e.g., two nearly identical electrical appliances without any distinguishing visual characteristic). We also removed instances where the semantic category of the target synset represents an abstract concept (e.g., "year" may be ambiguous to disambiguate through visual features). If after removing the synsets there were less than ten candidate synsets, we marked the instance to be augmented. After completing the test set manual validation, we augmented all instances with less than ten candidate synsets with random synsets selected from the other instances. Furthermore, we also manually created the query. To create the query, we combined: the lemma of the correct synset (the one we iterated over), one of the lemmas of the hypernym of the correct synset, and a word from the gloss of the correct synset. Using this methodology, we are sure that there is enough information to properly disambiguate the correct image. Furthermore, we randomize the order of each part of the query, in order to avoid participants being able to reverse engineer the construction of the query and obtaining information they should not have access to (without randomizing the order the first word of the query would always be the lemma of the correct synset). The word from the gloss should help disambiguate fine-grained semantics, while the lemma from the hypernym the high-level sense of the input. To increase the degree of diversity of the test split with respect to the train split, we also performed additional steps. We included synsets where glosses in Italian are not present. When using that synset as target, we manually translated the gloss from the original English one to Italian and selected the word to be used in the query. Furthermore, for all target synsets, we manually replaced the image from BabelNet with a different image sourced from sites with a permissive license.

For the train set, we do not provide a query directly. While creating the query using the lemmas of the target synset and its hypernym is easy, selecting a meaningful word from the gloss associated to the synset is not trivial. Participants are asked to develop a strategy to extract the word from the gloss to help with disambiguation. This also allows participants to have more freedom to use the dataset as they see fit, as long as the final model is capable of performing VWSD given only a query and a candidate pool of images.

4.2. Dataset Statistics

Listing 1: Example of Train Instance

```
 1  {
 2      "id": "bn:00022412n",
 3      "hyp_id": "bn:00017670n",
 4      "gloss": "Atto del cuocere",
 5      "lemma": "cucina",
 6      "hyp_lemma": [
 7        "cambiamento di stato"
 8      ],
 9      "bns": [
10        "bn:00018237n", ..., "bn:00049248n"
11      ],
12      "is_co_hyp": [
13        true, ..., false
14      ],
15      "images": [
16        "F14/bn:00018237n", ..., "F0/bn:00049248n"
17      ],
18      "all_lemmas": [
19        ["masticazione", "masticare"],
20          ...,
21        ["cucine", "cucina", "cucina attrezzata", "cucina aperta", "cucinotto"]
22      ],
23      "all_glosses": [
24        "La masticazione è il processo mediante il quale il cibo è frantumato e preparato
              dai denti.",
25        ...,
26        "Una stanza attrezzata per la preparazione dei cibi."
27      ],
28      "img": "F22/bn:00022412n"
29  }
```

For the train split, we provide a total of 10,000 instances. These instances are not formatted for the task, we provide the data of interest from BabelNet directly. Specifically, we provide for each instance the following fields:

- **'id'**: Synset id
- **'hyp_id'**: Synset id for the hypernym
- **'gloss'**: Gloss in Italian
- **'lemma'**: Lemma in Italian
- **'hyp_lemma'**: All possible lemmas for the hypernym
- **'bns'**: The list will contain both: co-hyponyms, synsets that have the "lemma" as a possible lemma
- **'is_co_hyp'**: True if the corresponding synset in "bns" is a co-hyponym, False otherwise
- **'images'**: Path to the image for each synset in "bns"
- **'all_lemmas'**: Each list contains all the Italian lemmas associated to each synset in "bns"
- **'all_glosses'**: All Italian glosses associated to each synset in "bns"
- **'img'**: Path to the image

Each instance is a unique synset in BabelNet. Participants are asked to format the dataset leveraging the provided data. We decided to implement this approach to let participants have complete freedom over the data and how they wanted to format the train set. For example, using this dataset to train

Listing 2: Example of Test Instance

```
1  {
2      "query": "patologia calcolo organo",
3      "candidates": [
4        "1133.jpg",
5        "850.jpg",
6        "743.jpg",
7        "1266.jpg",
8        "1367.jpg",
9        "948.jpg",
10       "549.jpg",
11       "695.jpg",
12       "1504.jpg",
13       "1148.jpg"
14     ],
15 }
```

a CLIP model (vision-language encoder) would require a different formatting with respect to a LLM supporting multimodal inputs. A complete example of train instance is reported in Listing 1.

For the test split, we provide a total of 222 instances, following the methodology previously described. Furthermore, image paths have been anonymized (since in the train split the image path was matched to the synset id of the instance) in order to avoid participants being able to extract additional information from Babel synsets at test time. A complete example of test instance is reported in Listing 2.

For both the train and test splits, all images are provided in a square size of 336. This is done in order to provide participants with the same experimental setting for visual inputs. Furthermore, this allows easier use of Large Language Models supporting multimodal inputs, since they treat visual inputs as sequences of tokens, smaller images are processed more efficiently with respect to higher resolutions.

5. Evaluation

5.1. Metrics

We will evaluate the models using the two metrics considered in the original challenge: HIT@1 and MRR. Given $r = [r_1, ..., r_n]$, where n is the cardinality of the test set and r_i is the rank of the correct image given as output by the model, MRR is defined as:

$$MRR = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{r_i} \tag{1}$$

This metric is used to evaluate the goodness of the ranking; the closer r_i is to 1 (i.e., the first position in the ranking), the better the result. HIT@1 is defined as:

$$HIT@1 = \frac{1}{n} \sum_{i=1}^{n} I(r_i) \tag{2}$$

where I is a function that returns 1 if $r_i == 1$ (i.e., the correct image is ranked first), and 0 otherwise. Therefore, this metric assesses the model's ability to select the correct image as the best possible candidate.

5.2. Baseline

We use the same baseline that was used in VWSD for SemEval 2023, that is a multilingual CLIP model[1] trained using the sentence-transformers library [19]. We do not perform additional fine-tuning of the model, we use the pre-trained model directly. We extract the embedding for the query text in the multimodal embedding space of the multilingual CLIP model, then we extract the embeddings for all candidate images for that query. We compute the cosine similarity for the query embedding with respect to all image embeddings and rank the candidate images in descending order with respect to cosine similarity (image with highest similarity at the top of the list). After obtaining the ranked list, we can compute HIT@1 and MRR. This method provides us with a strong baseline that doesn't include additional information from external sources or ensemble models.

5.3. Participants

One team - UniTor (University of Tor Vergata) - participated in the task and submitted three distinct runs. In this section, we briefly describe the system proposed by said participants.

5.3.1. UniTor

The team proposed a five-stage pipeline: semantic analysis and weighted binary question generation, visual assessment, visual question-answering confidence estimation, multimodal semantic matching, multi-channel fusion and ranking. The final scoring used in the ranking is computed using different scores obtained in the pipeline. Specifically, the scoring combines a visual-question answering score and a semantic similarity score obtained from a CLIP model. Semantic analysis and question generation is done by using a LLM (GPT-5.1). The visual-question answering score is obtained by leveraging the probability of generating a "Yes/No" answers of a smaller LLM (Qwen3-VL), while the similarity score is obtained by leveraging a textual description w.r.t. candidate image. We define the scoring methodology to report the results as follows:

- Score VQA: visual question-answering score obtained by leveraging the probability of generating a "Yes/No" answer;
- Score VQA (Two-Step): visual question-answering score obtained by leveraging the probability of generating a "Yes/No" answer. The questions for this score are generated using a more strict parsing approach;
- Score CLIP (Query): similarity score obtained by a CLIP model using the candidate image and the original query;
- Score CLIP (Description): similarity score obtained by a CLIP models using the candidate image and the description generated by a LLM.

Complete details regarding the participants solutions are detailed in their report [20].

6. Results

We report results for HIT@1 in Table 1 and for MRR in Table 2. Overall the second system proposed by UniTor performed better than all others in both the HIT@1 and MRR leaderboards. Still, the first proposed system performed remarkably well in terms of MRR (.8100 against the .8182 of the best system). This highlights that the first system is still capable of providing good rankings with respect to the best system. Furthermore, we highlight that all proposed systems by UniTor significantly outperformed the baseline. This highlights that their proposed solution was capable of overcoming a system based on pure semantic similarity scoring. Finally, the system proposed by UniTor leveraged Large Language Models supporting multimodal inputs. This highlights current interest in multimodal research, where researchers are not strictly interested in extraction of embeddings from vision-language

---

[1]https://huggingface.co/sentence-transformers/clip-ViT-B-32-multilingual-v1

encoder models, but are interested in leveraging Large Language Models supporting multimodal inputs for task resolution.

Team	Model	Score
baseline	clip-ViT-B-32-multilingual-v1	.4505
UniTor	Score VQA + Score CLIP (Query)	.6937
UniTor	Score VQA + Score CLIP (Query) + Score CLIP (Description)	**.7117**
UniTor	Score VQA (Two-Step) + Score CLIP (Query) + Score CLIP (Description)	.6667

Table 1
HIT@1 Leaderboard

Team	Model	Score
baseline	clip-ViT-B-32-multilingual-v1	.6244
UniTor	Score VQA + CLIP (Query)	.8100
UniTor	Score VQA + CLIP (Query) + CLIP (Description)	**.8182**
UniTor	Score VQA (Two-Step) + CLIP (Query) + CLIP (Description)	.7770

Table 2
MRR Leaderboard

7. Conclusions

In EVWSD-ITA, we proposed the first dataset for VWSD for fine-grained and high-level semantics in Italian. Given a query containing a target word and additional context, we combine both images representing the same sense as the target word (high-level semantics) and hard negatives obtained by exploiting the co-Hyponyms relation (fine-grained semantics). We provide an extensive train set obtained from BabelNet and a manually annotated test set. The EVWSD-ITA task was approached by a single team that proposed a solution based on Visual Question-Answering and semantic similarity obtained from a CLIP model. All participants' solutions were able to surpass the proposed baseline, for both metrics considered in this task (HIT@1 and MRR), highlighting the impact of their scoring strategy in VWSD.

Acknowledgments

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] R. Navigli, Word sense disambiguation: A survey, ACM computing surveys (CSUR) 41 (2009) 1–69.
[2] M. Bevilacqua, T. Pasini, A. Raganato, R. Navigli, Recent trends in word sense disambiguation: A survey, in: International joint conference on artificial intelligence, International Joint Conference on Artificial Intelligence, Inc, 2021, pp. 4330–4338.

[3] D. Sumanathilaka, N. Micallef, J. Hough, Assessing gpt's potential for word sense disambiguation: A quantitative evaluation on prompt engineering techniques, in: 2024 IEEE 15th Control and System Graduate Research Colloquium (ICSGRC), IEEE, 2024, pp. 204–209.

[4] A. Raganato, I. Calixto, A. Ushio, J. Camacho-Collados, M. T. Pilehvar, SemEval-2023 task 1: Visual word sense disambiguation, in: A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2227–2234. URL: https://aclanthology.org/2023.semeval-1.308/. doi:10.18653/v1/2023.semeval-1.308.

[5] F. Cutugno, A. Miaschi, A. P. Aprosio, G. Rambelli, L. Siciliani, M. A. Stranisci, Evalita 2026: Overview of the 9th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.

[6] T. Pasini, A. Raganato, R. Navigli, Xl-wsd: An extra-large and cross-lingual evaluation framework for word sense disambiguation, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 13648–13656.

[7] L. Huang, C. Sun, X. Qiu, X.-J. Huang, Glossbert: Bert for word sense disambiguation with gloss knowledge, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 3509–3514.

[8] J. H. Yae, N. C. Skelly, N. C. Ranly, P. M. LaCasse, Leveraging large language models for word sense disambiguation, Neural Computing and Applications 37 (2025) 4093–4110.

[9] D. Meconi, S. Stirpe, F. Martelli, L. Lavalle, R. Navigli, Do large language models understand word senses?, in: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025, pp. 33885–33904.

[10] A. Kritharoula, M. Lymperaiou, G. Stamou, Large language models and multimodal retrieval for visual word sense disambiguation, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 13053–13077. URL: https://aclanthology.org/2023.emnlp-main.807/. doi:10.18653/v1/2023.emnlp-main.807.

[11] S. Kwon, R. Garodia, M. Lee, Z. Yang, H. Yu, Vision meets definitions: Unsupervised visual word sense disambiguation incorporating gloss information, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1583–1598. URL: https://aclanthology.org/2023.acl-long.88/. doi:10.18653/v1/2023.acl-long.88.

[12] X. Zhang, T. Zhen, J. Zhang, Y. Wang, S. Liu, Srcb at semeval-2023 task 1: Prompt based and cross-modal retrieval enhanced visual word sense disambiguation, in: Proceedings of the 17th international workshop on semantic evaluation (SemEval-2023), 2023, pp. 439–446.

[13] Q. Yang, X. Wang, Y. Li, L.-K. Lee, F. L. Wang, T. Hao, Mta: A lightweight multilingual text alignment model for cross-language visual word sense disambiguation, in: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2024, pp. 12166–12170.

[14] Y. Laba, Y. Mohytych, I. Rohulia, H. Kyryleyza, H. Dydyk-Meush, O. Dobosevych, R. Hryniv, Ukrainian visual word sense disambiguation benchmark, in: Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP)@ LREC-COLING 2024, 2024, pp. 61–66.

[15] I. Setitra, P. Rajapaksha, A. K. Myat, N. Crespi, Leveraging ensemble deep models and llm for visual polysemy and word sense disambiguation, Multimedia Tools and Applications (2025) 1–33.

[16] E. Musacchio, L. Siciliani, P. Basile, G. Semeraro, Assessing and improving the multilingual visual word sense disambiguation ability of vision-language models, in: ECAI 2025, IOS Press, 2025, pp. 4145 – 4152.

[17] J. Robinson, C.-Y. Chuang, S. Sra, S. Jegelka, Contrastive learning with hard negative samples, arXiv preprint arXiv:2010.04592 (2020).

[18] R. Navigli, S. P. Ponzetto, Babelnet: Building a very large multilingual semantic network, in: Proceedings of the 48th annual meeting of the association for computational linguistics, 2010, pp. 216–225.

[19] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: http://arxiv.org/abs/1908.10084.

[20] C. D. Hromei, A. Scaiella, D. Croce, R. Basili, Unitor at evwsd-ita: Zero-shot visual word sense disambiguation via visual question-answering, in: Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), CEUR.org, Bari, Italy, 2026.